

'HOT SPOT' ANALYSIS USING BOTH THE SYSTAT 'K-MEANS' ROUTINE AND A RISK ASSESSMENT

by

Ned Levine
Ned Levine & Associates
Annandale, VA

In this report, I will discuss two different approaches to 'hot spot' analysis. The first is the use of the Systat 'K-Means' routine to conduct a partitioning cluster analysis of k pre-specified clusters. The second is an attempt to reduce the overwhelming effect of population density on 'hot spot' estimation through a risk assessment.

Partitioning Cluster Analysis

Cluster analysis is a technique for grouping observations together based on similarities on one or more variables. The observations in a cluster (or group) are assumed to be more homogenous amongst themselves than they are with observations that are not in the cluster. However, there are several different families of cluster analysis and many variations within each family, any of which will produce unique groupings (Everitt, 1974; Systat, Inc. 1994).

The two most common of these families are the hierarchical and partitioning clustering methods. For example, hierarchical clustering groups incident locations together based on spatial proximity (Everitt, 1974; Systat, Inc. 1994). In pure hierarchical clustering, the technique will first group all observations into pairs, based on the shortest distance between points (nearest neighbors). It will then group all pairs into quartets, again based on the shortest distance between the centroid of the pairs. The quartets will be grouped into octets and so forth until all points converge into a single cluster. The user can then select which of the levels of clustering is desired, usually based on an F-test of between-to-within group variance.

On the other hand, partitioning techniques group observations together based on a pre-specified number of clusters. The user defines the number of clusters into which the data are to be separated, k , and the variables used to group cases into clusters. For spatial methods, the variables would be, of course, the X-coordinate and the Y-coordinate combined with any variable that might be used for weighting cases (e.g., population). The algorithms proceed to allocate all points to one, and only one, of these k groups. However, different partitioning routines differ on how observations are grouped together - the choice of a grouping variable, the measurement of distance, the initial starting point (seed), the assignment rule, and the stopping criterion. Thus, even within families, techniques will produce different solutions.

Systat 'K-Means'

For example, both the *Systat* 'K-means' routine and the SAS 'Proc Fastclus' routine implement partitioning techniques in which the data are placed into k discrete groups, maximizing between-cluster variation relative to within-cluster variation. However, the two routines use different approaches that produce varying results. The *Systat* 'K-Means' routine starts with one seed (the mean center) and splits the data into two clusters by picking the case farthest from the center as a seed for a second cluster. It then assigns each case to the nearest cluster center and recalculates the mean X and mean Y of these two clusters (the cluster means). These means are the seeds for the next splitting (partitioning) of the data. For the third seed, the program chooses the data point whose combined distance from the first two cluster means is greatest. The program then reassigns all points to the nearest of the three clusters, then recalculates the means of each cluster. It continues this process creating a fourth, a fifth, and so forth only stopping when the number of clusters reaches k , the initially specified number. The program then assigns all cases to one, and only, cluster and calculates an F-test of between group to within group variance of distances.

On the other hand, SAS 'Proc Fastclus' starts with guesses about where the k means will be (the first k observations that are separated by a certain minimum distance, either user specified or internal to the program). Then, it assigns each case to the initial mean to which it is closest and re-calculates the cluster means. It then repeats this process until changes are very small. The problem with this approach is that it does not produce consistent results. Because the routine defines initial seeds by selected cases (e.g., the first seed is the first observation in the data set), differing results can be obtained by selecting different initial cases. For example, when I sorted the data set first by latitude and then selected 10 clusters, SAS 'Proc Fastclus' gave me one set of results. However, when I then sorted the data set by longitude and selected 10 clusters, I got slightly different results. The routine is not consistent from one replication to another. Consequently, for this analysis, I decided only to use the *Systat* 'K-means' routine which does produce consistent results.

Program Use

The routine is easy to use. Since it is a general statistical package, *Systat* has a common command syntax and uses menu items to cover many statistics. The *K-means* routine is standard in both the DOS and Windows versions of the program and can be accessed either through a menu or commands. Within the program, the user specifies that the category of analysis is cluster analysis. Then, the user indicates that *K-means* is to be used and defines the grouping variables (latitude and longitude, in our case). Finally, the user specifies the number of clusters to be extracted. The program outputs a file that assigns each case (numbered from 1 to n) into one, and only one, of the k clusters and produces an F-test of between-group to within-group variance.

There are a number of statistical options. The first is to standardize the variables in order to assign equal weight. I didn't do this since the extent of Baltimore County is relatively even in the X- and Y-directions. But it would be important if the shape of the geographical area were exaggerated (e.g., long and thin). There are several different distance metrics allowed. I used Euclidean because we were approximating two-dimensional distances. But

there are distance measurement options for Gamma, Pearson, R-squared, Minkowski, Chi-square, Phi-square, and minimum within sum of squares deviations. It's not clear how the use of these different metrics would affect spatial clustering since I only used Euclidean.

The Number of Clusters Extracted

The key decision, however, is the number of clusters to be extracted. How does one know how many clusters to extract? It depends on a user's purpose. For example, if theory or prior experience says there should be a certain number of clusters, then that number should be used for the number of partitions. One statistical criteria that can be used is to examine the F-test of between-group to within-group variance. The number of partitions with the highest F-test would usually have the best differentiation, that is splitting cases into clusters so that between-group distances are maximized while within-group distances are minimized. Even here, however, there is not a correct number as generally the F-test will increase as the number of clusters extracted increases due primarily to increasing spatial differentiation in the data set. At the extreme, one would have maximum separation by partitioning the n cases into n clusters, a result that would not be very useful, however.

The Application of K-Means to Baltimore County Residential Burglaries and Street Robberies

The approach I used with the test data set was to examine the results for increasing numbers of partitions, using the F-test as a rough criteria. That is, for both 1996-97 residential burglaries and street robberies separately, I partitioned the data into two clusters. Then, I partitioned the data into three clusters. Then, I partitioned the data into four clusters, and so forth until I had divided the data into 35 separate partitions. I then examined the 34 F-tests to see which partitioning had the highest F-value.

For the 1215 'clean' geocoded robberies, this method worked smoothly and 30 partitions seemed like an optimal amount.¹ However, for the 6214 residential burglaries, the data set was very large. Neither version 5 or version 6 of *Systat* (both DOS programs) would construct a partitioning with the data set. With the Windows version 6.1, the program did construct clusters with this data set, but only up to 25 partitions. Consequently, I accepted the limits of the program and partitioned the residential burglaries into 25 clusters. For each cluster, I used the *CrimeStat* ellipse algorithm to calculate the standard deviational ellipse and outputted the result to *Atlas*GIS* (Levine, 1996; Levine and Canter, 1998). Since the ellipse is similar to a bidirectional standard deviation, approximately two-thirds of the cases are found within each ellipse; the exact proportion depends on dispersion and skewness in the distribution, however.

Figure 1 shows the location of 1996-97 street robberies in Baltimore County. Figure 2 shows the distribution of 22 extracted clusters from the K-means routine. Eight of the clusters had fewer than 5 cases, too few to produce a standard deviational ellipse. At face value, the

1. Even though the initial data set had 1252 street robberies and 6467 residential burglaries, there were 14 duplicate records and 276 observations which had one or both coordinates missing. Consequently, the data set used had 1215 robberies and 6214 burglaries.

Figure 1: 1996-97 ROBBERIES IN BALTIMORE COUNTY

Location of Street Robberies

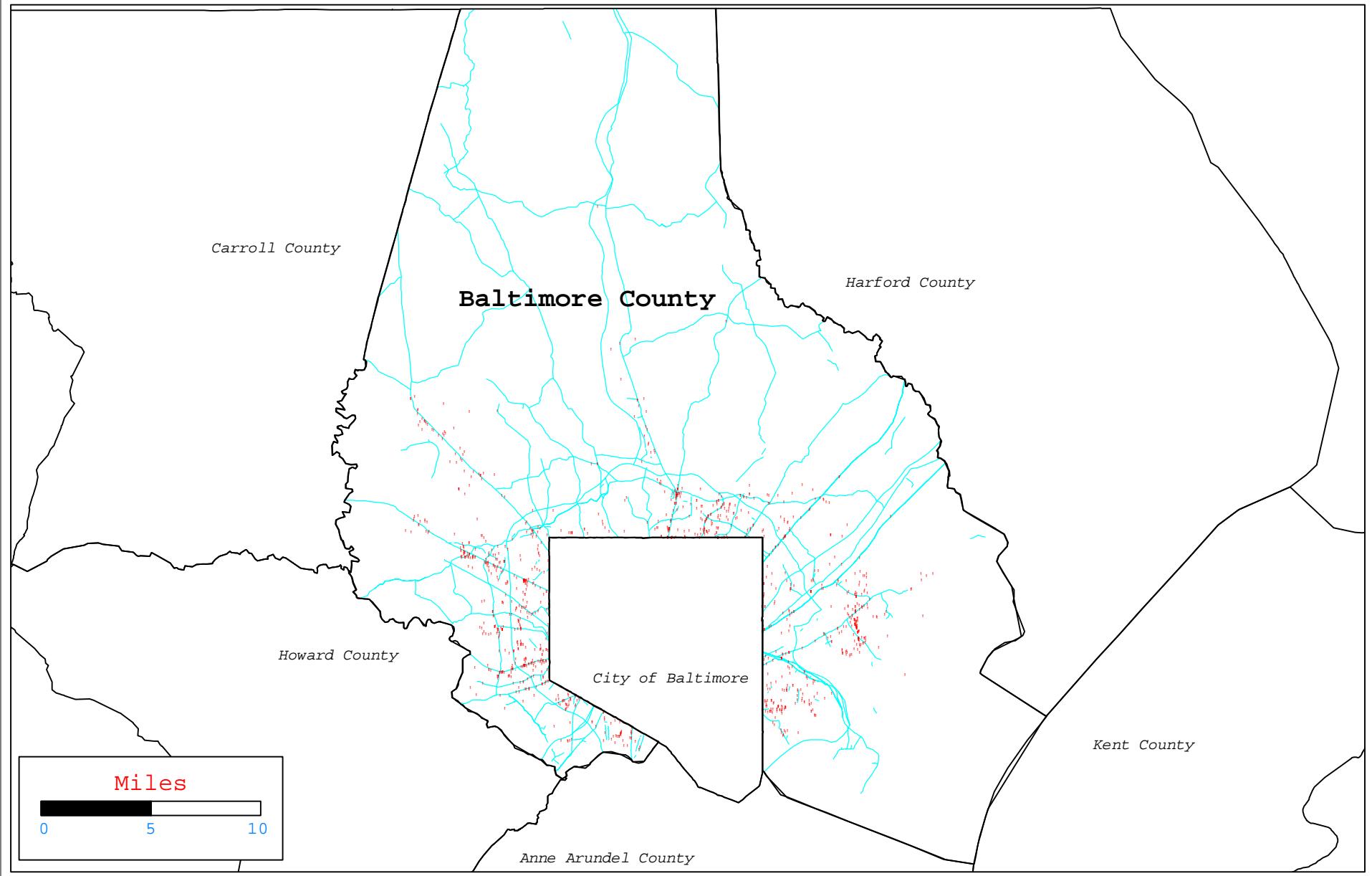
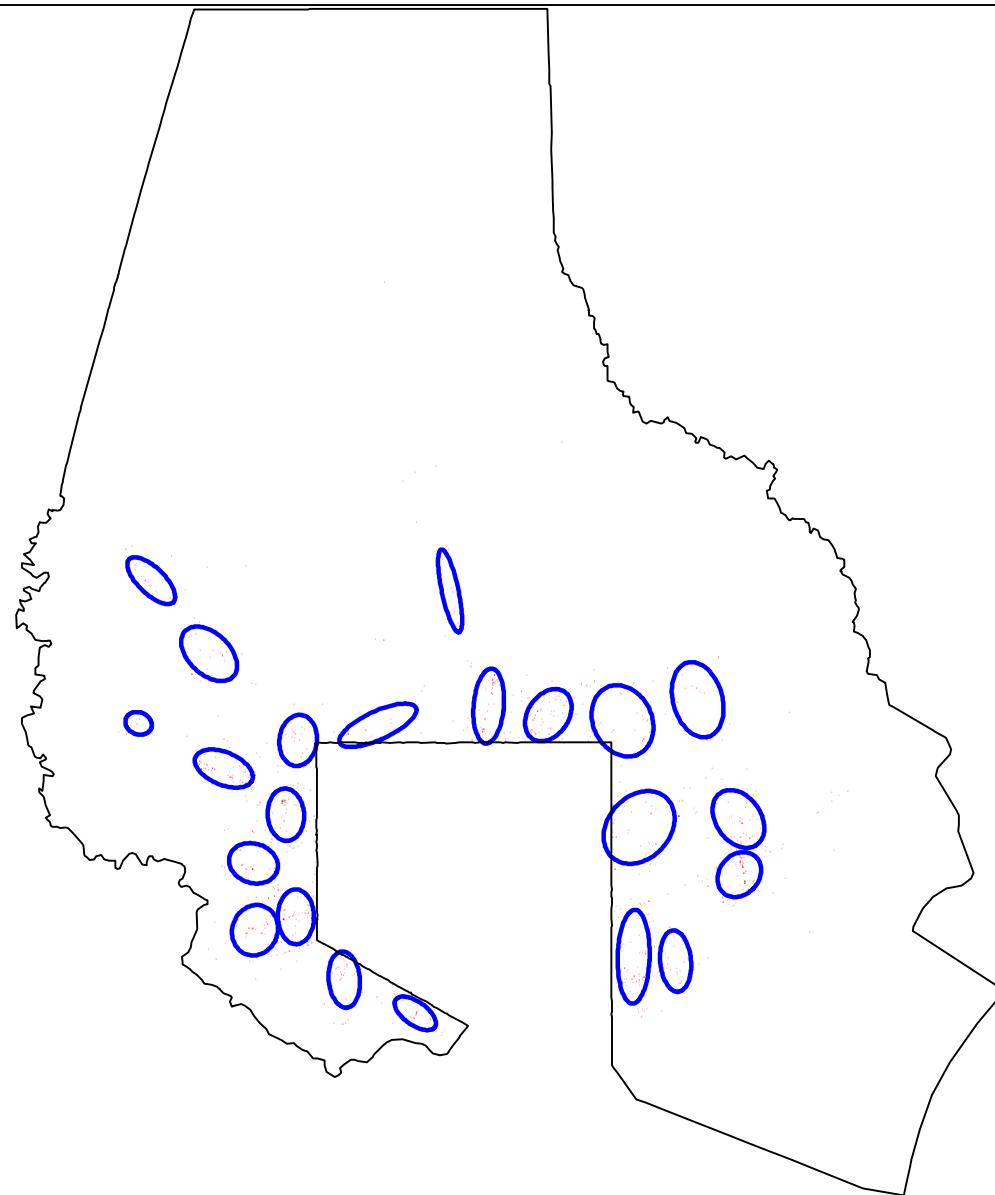


Figure 2: K-MEANS ANALYSIS OF ROBBERY 'HOT SPOTS'

22 of 30 Extracted Clusters



distribution of the ellipses seems reasonable, particularly nearer the border with Baltimore City. For a couple of clusters, the results seem forced. Figure 3 shows a blow-up of the southwest area of Baltimore County. What is interesting is the concentration of the clusters near the border with Baltimore City, the result of higher population and employment concentration.

Figure 4 shows the location of residential burglaries in Baltimore County. Figure 5 shows the results of partitioning the burglaries into 25 clusters and Figure 6 shows a blow-up of southwest Baltimore County. Unlike the robberies, there are clusters located throughout the county. As with robberies, some of these seem reasonable while others seemed forced. Finally, figure 7 shows the overlap in the ellipses between street robberies and residential burglaries. There are a number of overlapping ellipses, particularly near the border with Baltimore City.

Summary of K-Means Analysis

The use of a *K-means* is appropriate if the user takes the time to determine an optimal number of clusters to extract. Because it requires the user to input the number of partitions, there is no automatic solution to a clustering problem. The user must experiment with different numbers of clusters. On the other hand, this approach has methodological soundness in that it will get the user to think about whether clusters are meaningful or not. The program is easy to use. I experimented with different solutions by adding other variables (e.g., population size). In the end, I stuck with using just latitude and longitude as my grouping variables.

Alternative Risk-Based Approach

One of the problems with *K-Means* and other clustering techniques is that they mix up first-order with second-order statistical results (Bailey and Gatrell, 1995). Spatial statisticians refer to *first-order effects* as indicating the general spatial trend within an area while *second-order effects* are the local spatial arrangements. With most metropolitan areas in the United States, much of the clustering is due to higher concentrations of population towards the urban center. In the case of Baltimore County, the higher concentration of population within Baltimore City means that, all other things being equal, clusters of cases will fall along the border with the city. This dominant pattern disguises any local effects as the distance between incident locations will necessarily decrease as one approaches the urban center.

To see this, we constructed some additional analysis. Figure 8 shows a map of population density by 1990 census block group for the Baltimore metropolitan region. As with almost all metropolitan areas everywhere, there is a much higher concentration of population in the central city. Figure 9 shows the approximate population density of block groups as a function of distance from an arbitrarily chosen point in the Baltimore harbor. This is called a *distance density function*. Had a different origin been taken, the distribution would simply be shifted slightly to the right or left. To smooth the distribution, a moving average of 15 proximal observations has been taken. As seen, there are very high concentrations of population in the central core. The density is low around the harbor, which is primarily a commercial/office area, but rapidly rises to a peak of about 35,000 persons per square mile. It then drops off quite rapidly. The borders of Baltimore County vary from 3.3 miles to 8.3 miles away from the origin, with an average of 5.8 miles. Population densities at this range vary from about 15,000

Figure 3: K-MEANS ROBBERY 'HOT SPOTS'

Southwest Baltimore County

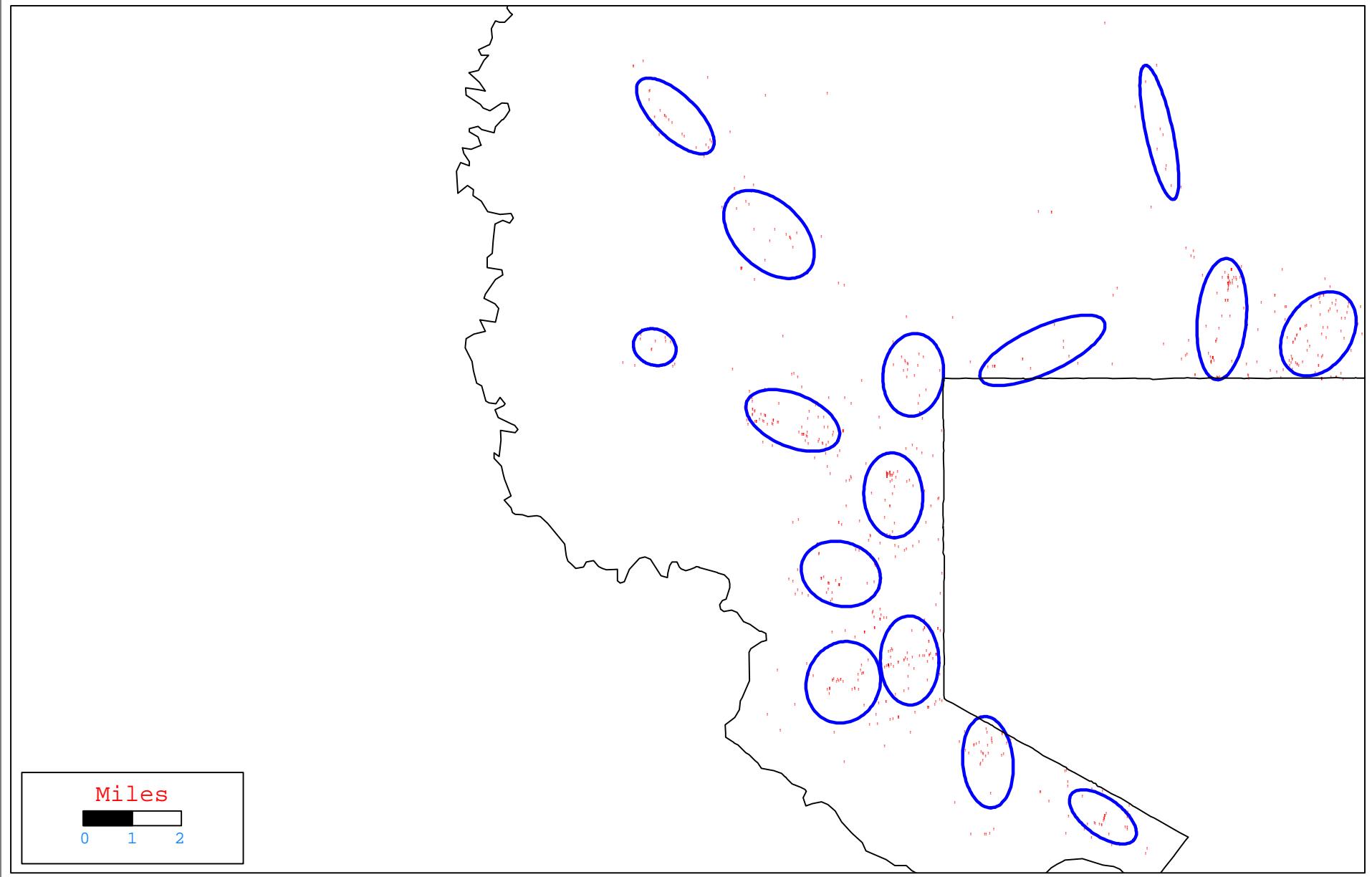


Figure 4: 1996-97 BURGLARIES IN BALTIMORE COUNTY

Location of Burglaries

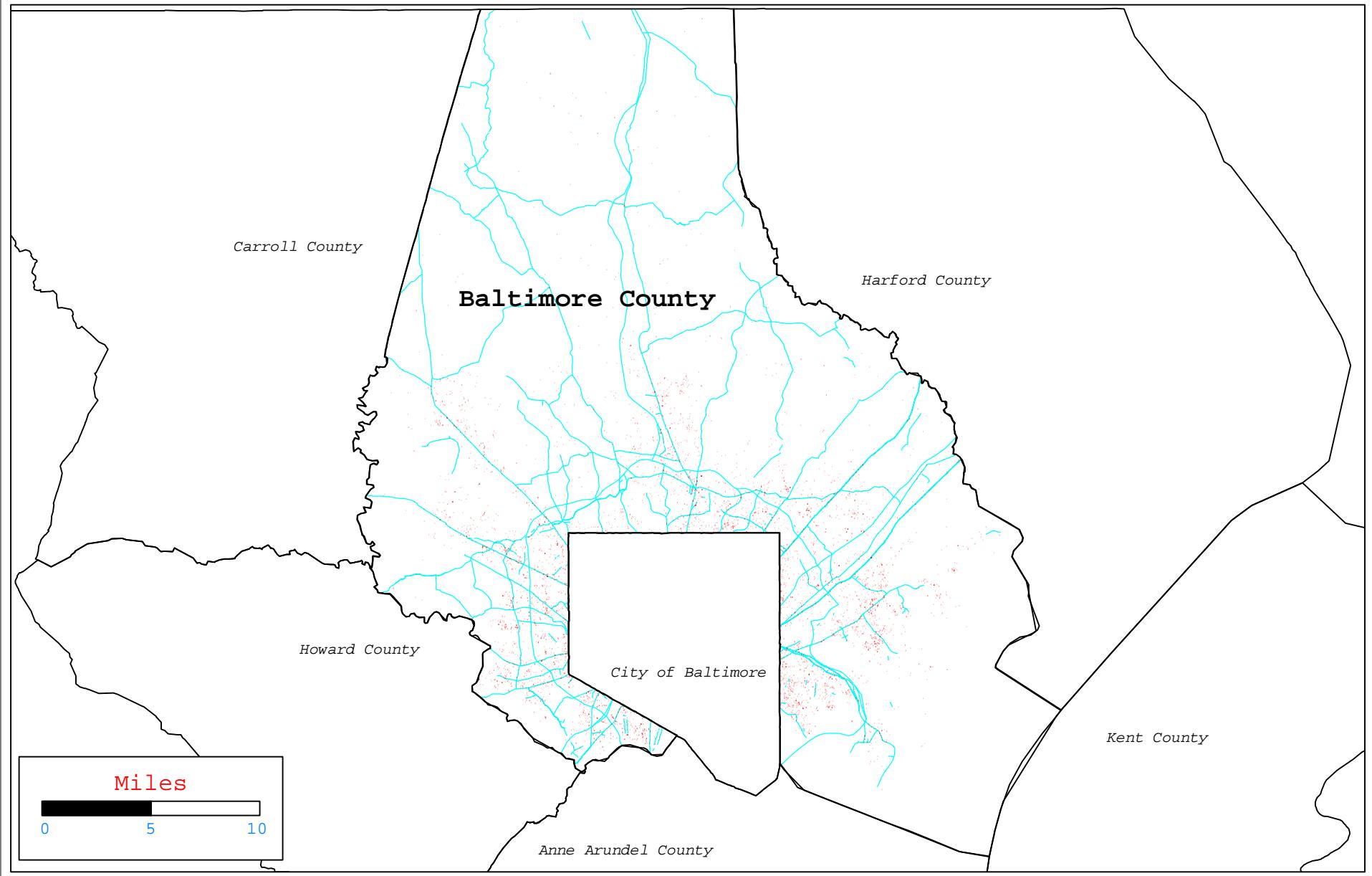


Figure 5: K-MEANS ANALYSIS OF BURGLARY 'HOT SPOTS'

25 Clusters

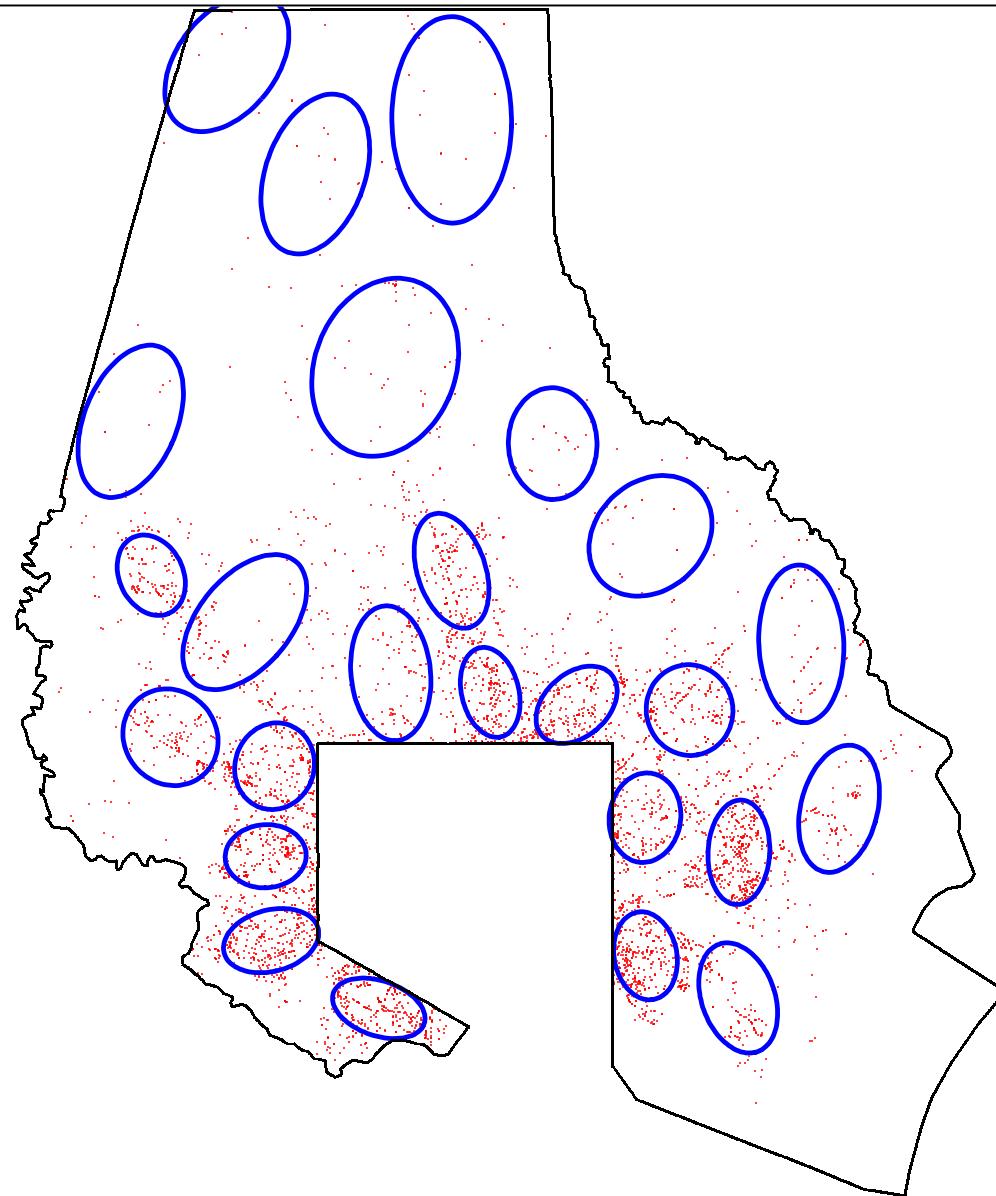


Figure 6: K-MEANS BURGLARY 'HOT SPOTS'

Southwest Baltimore County

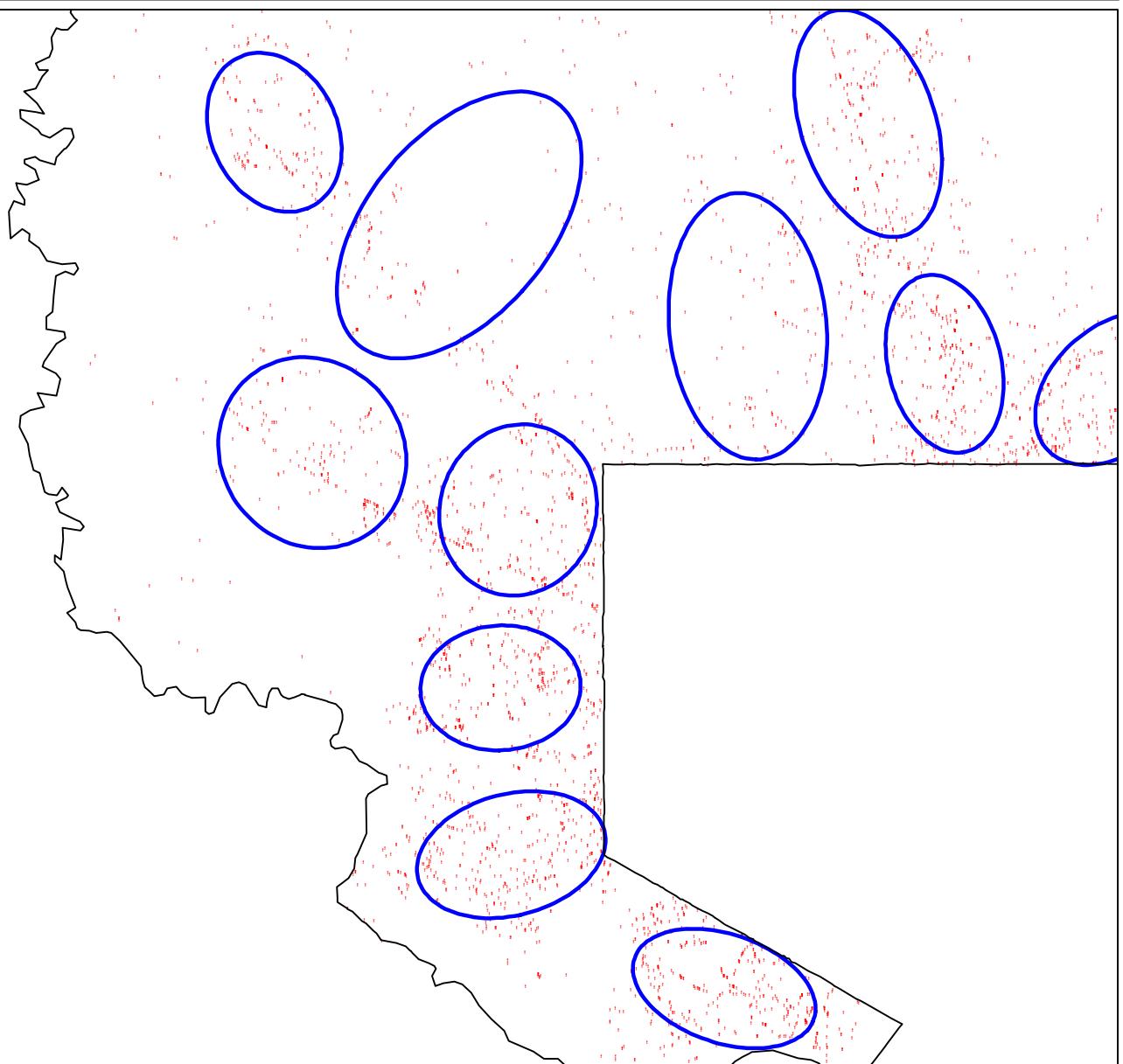


Figure 7: K-MEANS ANALYSIS OF 'HOT SPOTS'

Overlap of Robbery and Burglary Clusters

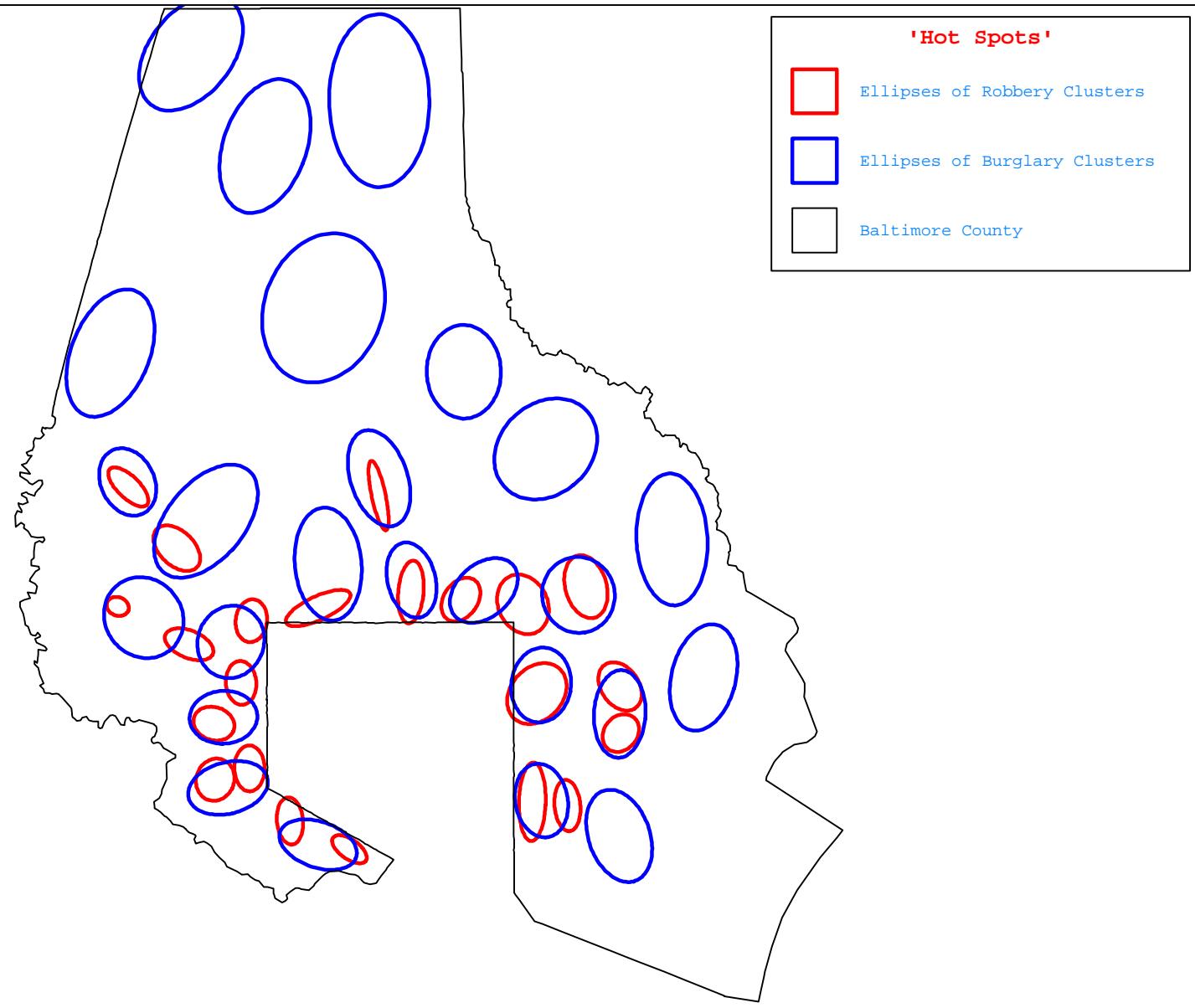


Figure 8: 1990 POPULATION DENSITY IN THE BALTIMORE REGION

Persons Per Square Mile

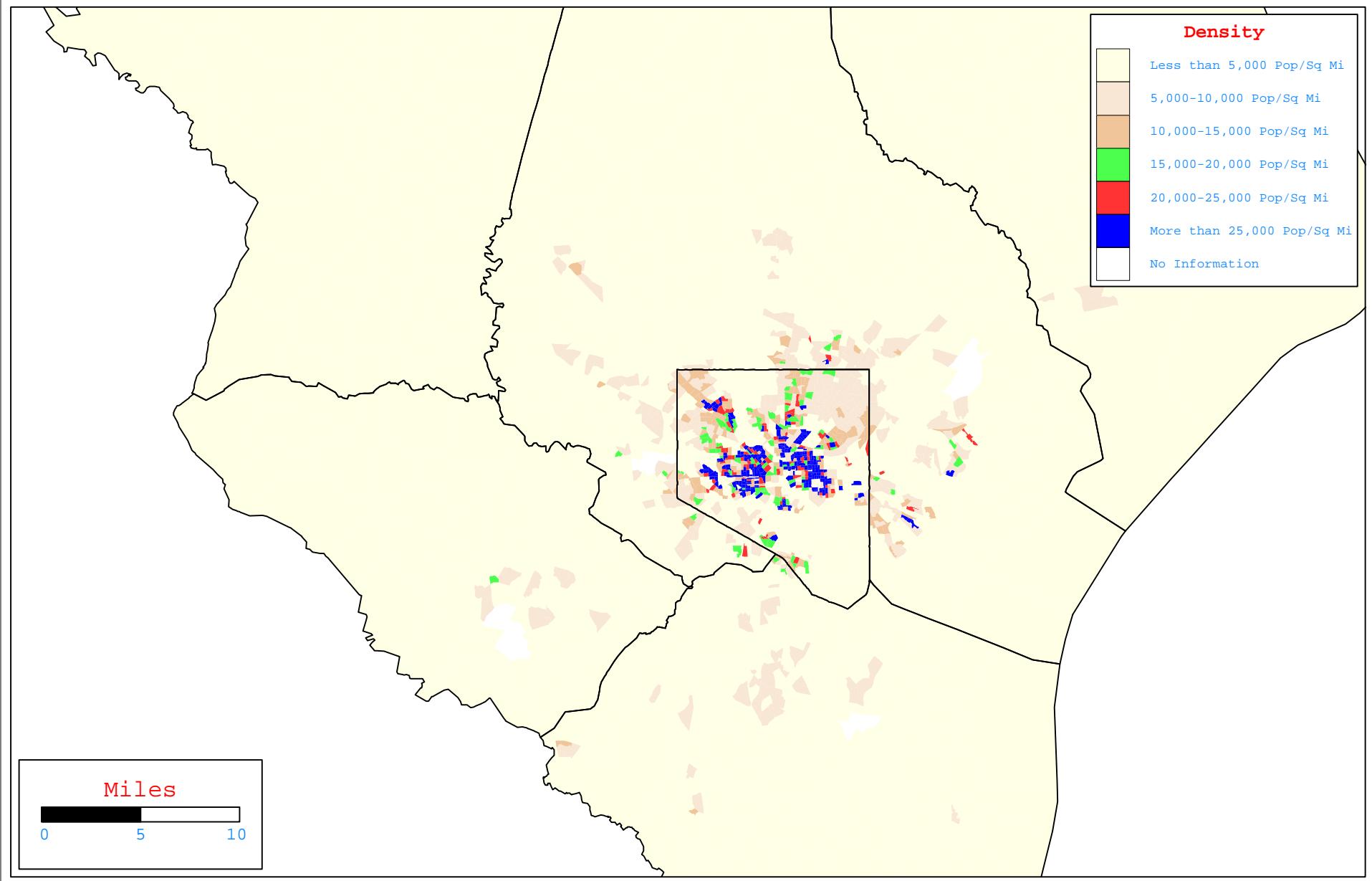
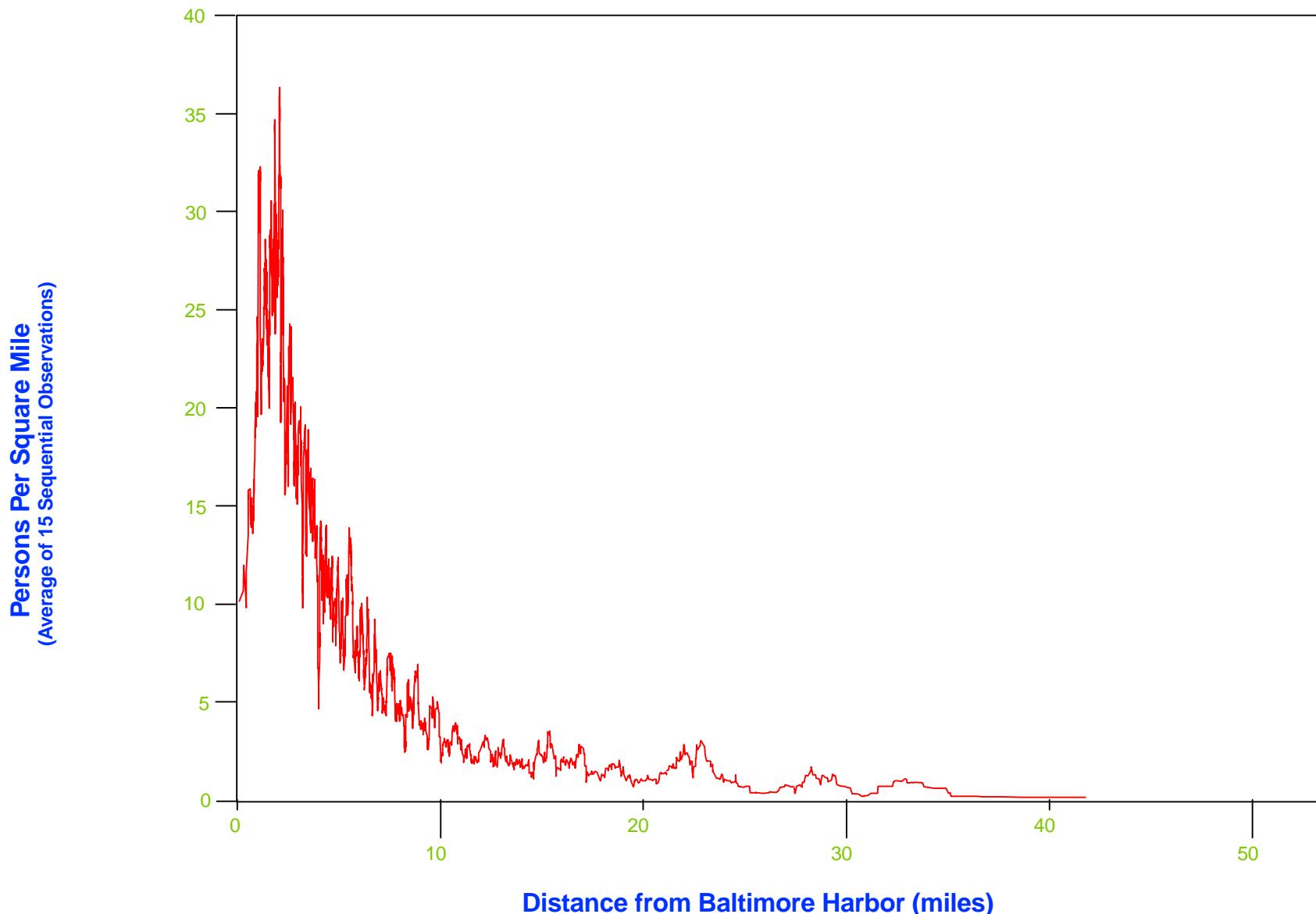


Figure 9:
POPULATION DENSITY IN THE BALTIMORE REGION: 1990
Moving Average of 15 Observations



persons per square mile to about 5,000 persons per square mile. Densities drop off rapidly away from the border with Baltimore City. In other words, the highest concentration of crime incidents in Baltimore County will necessarily be along the border with Baltimore City because there is a higher concentration of population.

This illustrates the effect of density on crime risk. Higher concentrations of people lead to a higher number of incidents, all other things being equal. Of course, concentration of people will vary by time of day and day of week. During the nighttime, the residential population is a good approximation for population concentration. During the daytime, however, the distribution of employment - which is even more concentrated, might be an even better indicator. In our test, we selected population as a base because this data was easily available. However, for certain types of crimes, such as street robberies, employment might be a better choice since street robberies tend to concentrate in commercial areas where there is a large daytime population. This is also the reason why street robberies show a more concentrated pattern than residential burglaries.

For 'hot spot' analysis, the higher densities due to population or employment distribution predispose identified 'hot spots'. Any clustering method - *K-Means*, hierarchical nearest neighbor, STAC, kernel density estimates, will always produce a lot of clusters where there is a higher concentration of people. This doesn't indicate that these locations are particularly risky, only that they have higher numbers of people. Obviously, different users may find this useful or not. Police, for example, may find this an adequate method for identifying places to assign officers. Since the assignment of officers has to be done efficiently, allocating personnel where incidents are frequently occurring is administratively sound. From the viewpoint of understanding unique environments, however, having a high concentration of incidents due to a high concentration of people is not very enlightening and certainly doesn't contribute much to an understanding of how the urban environment creates crime centers. From a statistical viewpoint, using a method for identifying 'hot spots' based on concentrations of incidents is mixing up first-order with second-order effects.

Controlling for Population Size

To examine unique environments that have 'more than their share' of crime, we need to minimize the effects of population size. A more exact approach would be to create two different density surfaces, one for the number of crime incidents and one for population density (population divided by area). The ratio of the two surfaces would then be constructed and tested against a random distribution (Bailey and Gatrell, 1995, 126-128), For this exercise, I did not have the tools available to construct such a model.

Consequently, as a rough approximation, crime incidents were standardized by population size. Robberies were assigned to individual block groups and then summed for each block group. They were then divided by the 1990 population and further divided by 10,000 to produce an *approximate* robbery risk rate (robberies per 10,000 population). The rate is approximate because 1996-97 robberies are being divided by 1990 population; population shifts since 1990 may have altered the relative distribution of these rates slightly.

The distribution for street robberies is highly skewed. Most block groups have low robbery risk while a few have a high risk. The median is about 8.5 robberies per 10,000 population while the mean is about 38.7 robberies per 10,000 population. Only 14% of the block groups have a robbery rate higher than the mean. Consequently, the mean was used as an index of high robbery risk. Figure 10 shows a map of 1990 block groups that have above-average numbers of robberies per 10,000 population. As seen, they are distributed throughout the built-up part of Baltimore County, not just along the border.

With residential burglaries, a similar procedure was followed. However, two different baselines were taken - population and households. Figure 11 shows block groups having a *per capita* burglary rate greater than 156 burglaries per 10,000 population (the mean for all block groups) while figure 12 shows block groups having a *per household* burglary rate greater than 392 burglaries per 10,000 households (the block group mean). The two maps are very similar with only a few discrepancies. A comparison of these two maps with Figure 10 shows there are many areas that have both high robbery risk and high burglary risk. These would be good candidates for 'high crime hot spots'.

Summary of Risk Analysis

In other words, when population is used as a standardizing variable, certain areas are identified that have higher risks for robbery or burglary above-and-beyond the sheer volume of incidents, which is primarily a function of population size. This concept of 'hot spots' is closer to the second-order effect that statisticians discuss (Ripley, 1981; Bailey and Gatrell, 1995). These are places that have unique properties which generate crimes in excess of population concentrations. For street robberies, among the community factors that are probably correlated are the existence of commercial corridors, high traffic volumes, the existence of public housing, high poverty levels, and high unemployment. For residential burglaries, possible correlates are community wealth and nearby commercial areas. Further research could attempt to isolate particular land use characteristics associated with these areas.

Figure 10: 'HOT SPOT' ROBBERY ZONES IN BALTIMORE COUNTY
1990 Block Groups with Above-Average Robberies Per 10,000 Population

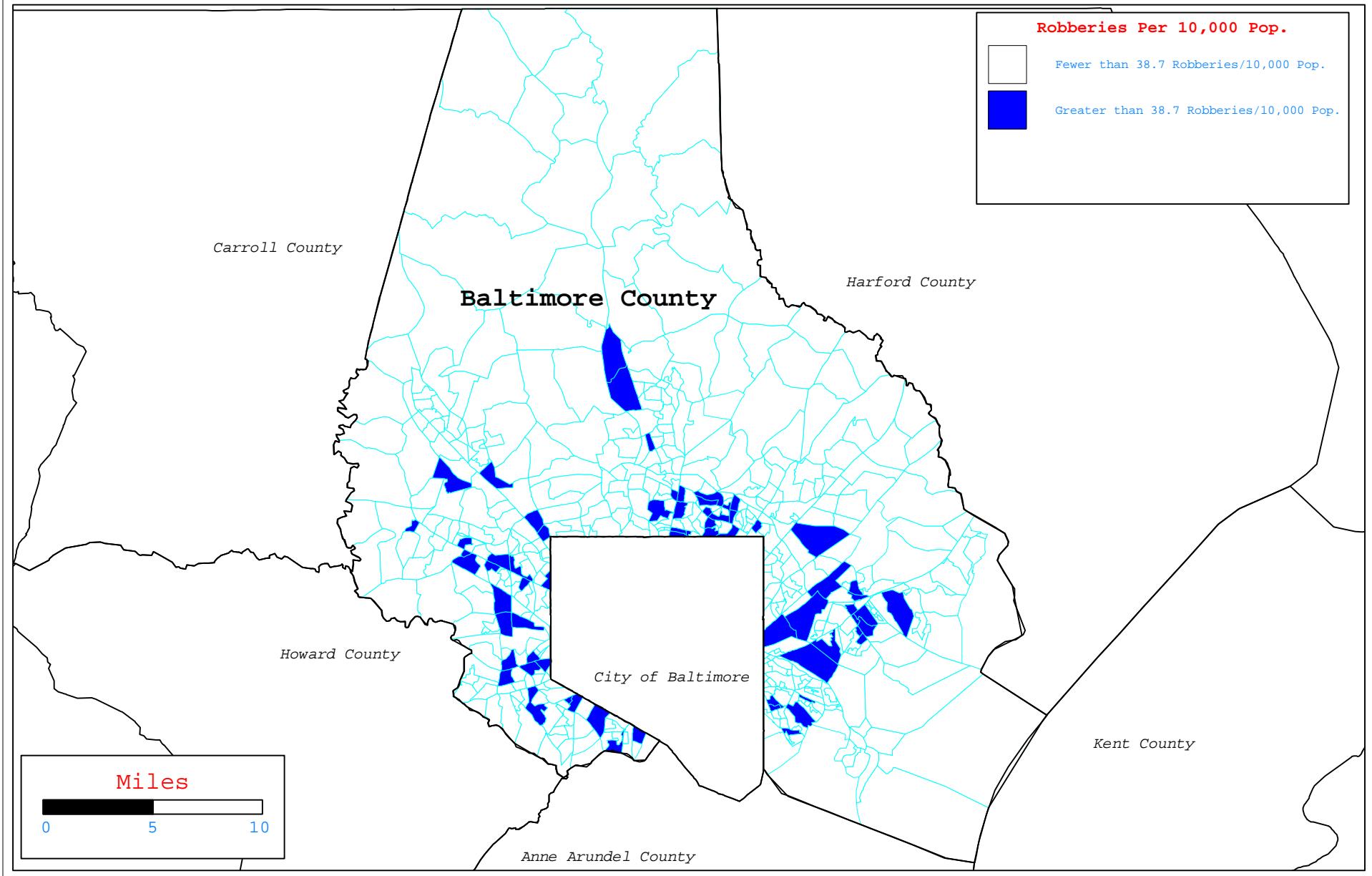


Figure 11: 'HOT SPOT' BURGLARY ZONES IN BALTIMORE COUNTY - I

1990 Block Groups with Above-Average Burglaries Per 10,000 Population

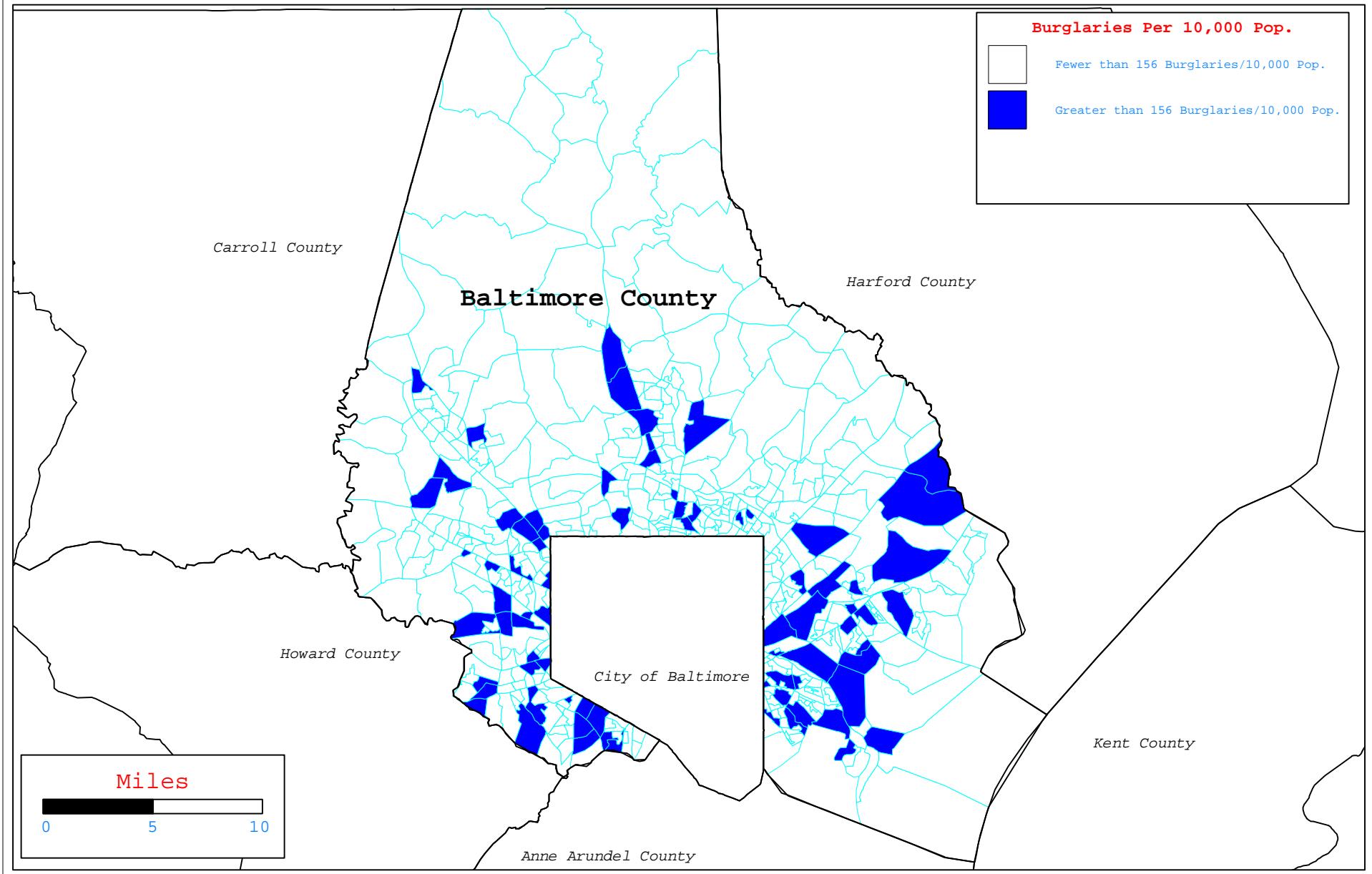
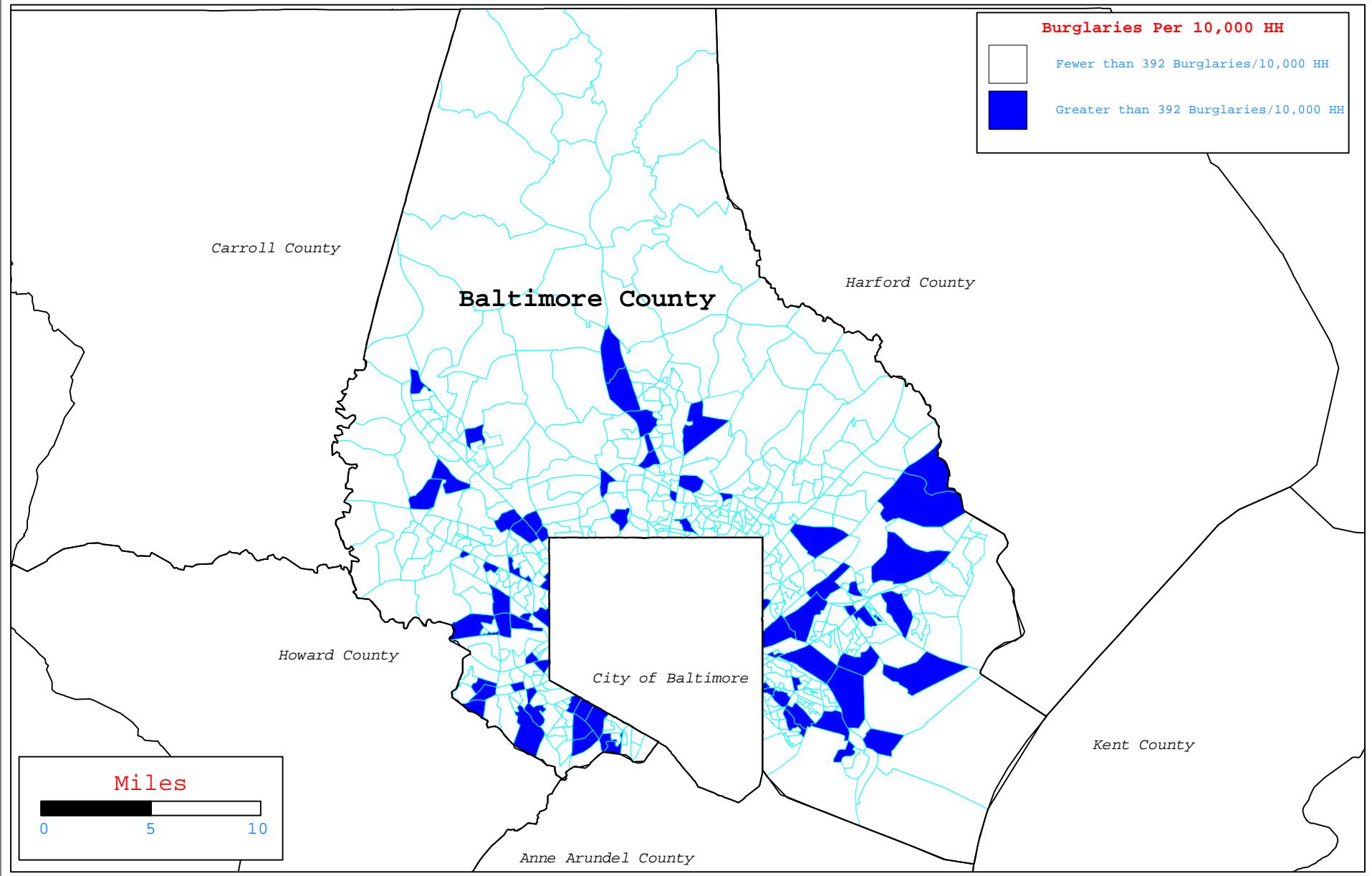


Figure 12: 'HOT SPOT' BURGLARY ZONES IN BALTIMORE COUNTY - II

1990 Block Groups with Above-Average Burglaries Per 10,000 Households



REFERENCES

- Bailey, Trevor C. and Anthony C. Gatrell (1995). *Interactive Spatial Data Analysis*. Longman Scientific & Technical: Burnt Mill, Essex, England.
- Everitt, Brian (1974). *Cluster Analysis*. Heinemann Educational Books: London.
- Levine, Ned (1996). "Spatial statistics and GIS: software tools to quantify spatial patterns". *Journal of the American Planning Association*. 62 (3), 381-392.
- Levine, Ned and Phil Canter (1998). *CrimeStat* - a spatial statistical program for crime analysis: a status report. NIJ Cluster Conference on the Development of Spatial Analysis Tools. February 26-27. Washington, DC.
- Ripley, B. D. (1981). *Spatial Statistics*. New York: John Wiley and Sons.
- Systat, Inc. (1994). *Systat for DOS: Advanced Applications, Version 6 Edition*. Systat, Inc.: Evanston, IL.